

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

Use of RNA Secondary Structure for Evolutionary Relationships:
Investigating RNase P and RNase MRP

A thesis presented in partial fulfilment of the requirements

For the degree of

Master of Science in Genetics

At Massey University

New Zealand

Lesley Joan Collins

1998

“Science teaches us about the deepest issues of origins, natures, and fates – of our species, of life, of our planet, of the Universe. For the first time in human history, we are able to secure a real understanding of some of these matters. Every culture on Earth has addressed such issues and valued their importance. All of us feel goosebumps when we approach these grand questions. In the long run, the greatest gift of science may be in teaching us, in ways no other human endeavour has been able, something about our cosmic context, about where, when, and who we are.”

Carl Sagan - The Demon-Haunted World.

Amendments:

Page 81: Top lines of page should read - The RNAstructure tree (Figure 5.2C) groups the chloroplast sequences together but does not group the cyanobacterial species (*Synechocystis*, *Anabaena* and *Anacystis*) together unless the *E. coli* outgroup is removed.

Page 126: The following reference should be included:

Pascual, A. and Vioque, A. (1996) Cloning, purification and characterisation of the protein subunit of ribonuclease P from the cyanobacterium *Synechocystis* sp. PCC 6803. *Eur J Biochem* 241: 17-24

Abstract

Bioinformatics is applied here to examine whether RNA secondary structure data can reflect distant evolutionary relationships. This is important when there is little confidence in sequence data such as when looking at the evolution of RNase MRP (MRP).

RNase P (P) and RNase MRP (MRP) are ribonucleoproteins (RNPs) that are involved in RNA processing and due to functional and secondary structure similarities, are thought to be evolutionary related. P activity is found in all cells, and fits the criteria for inclusion in the RNA world (Jeffares et al. 1998). MRP is found only in eukaryotes with essential functions in both the nucleus and mitochondria. The RNA components of P and MRP (pRNA and mrpRNA) cannot be aligned with any certainty, which leads to a lack of confidence in any phylogenetic trees constructed from them.

If MRP evolved from P only in eukaryotes then it is an exception to the general process of the transfer of catalytic activity from RNA, to ribonucleoproteins, to proteins (Jeffares et al. 1998). An alternative possibility that MRP evolved with P in the RNA world (and has since been lost from all but the eukaryotes) is raised and examined. Quantitative comparisons of the pRNA and mrpRNA biological secondary structures have found that the third possibility of an organellar origin of MRP is unlikely.

Results show that biological secondary structure can be used in the evaluation of an evolutionary relatedness between MRP and P and may be extended to other catalytic RNA molecules. Although there are many protein families, this may be the first evidence of the existence of a family of RNA molecules, although it would be a very small family.

Secondary structures derived with folding programs from pRNA and mrpRNA sequences are examined for use in the characterisation of catalytic RNA sequences. The high AT content in organellar genomes may hinder the identification of their catalytic RNA sequences. A search strategy is developed here to address this problem and is used to identify putative pRNA sequences in the chloroplast genomes of four green plants. A maize chloroplast pRNA-like sequence is examined in more detail and shows many characteristics seen in known pRNA sequences. Folding programs show some potential for the characterisation of possible catalytic RNA sequences with only a small bias in the results due to sequence length and AT content.

Acknowledgments

Many, many thanks must go to David Penny for his patience and long hours in the air reading this thesis. I appreciate all the work that has gone into getting this bench jockey to wonder into the (RNA) world of theoretical science.

Special thanks to Vince Moulton (the ever travelling mathematician) who was game to team up with the crazy biologist and put some 'real' data into DCA. Thanks also to Soeren Perrey for teaching me the basics of DCA and Unix (a language even stranger than Klingon). Thanks also goes to Robert Pointon for writing all of those wonderful little programs that made life so much easier.

Thanks must also go to the inhabitants of the Boffin lounge for support and occasional coffee. Thanks also to the inhabitants of the BN lab at the NZDRI, for the coffee and time, especially during the writing of this thesis. Many thanks to all my family for their support over the years.

A great, great many thanks must go to my husband Maurice whose unwavering support over the last few years has gotten me through this. I could not have done this without you.

Finally, to everybody who is adventurous enough to work on the fringe... **Qaplah!**

(Klingon for Success)

Table of Contents

Abstract	iii
Acknowledgments	iv
Table of Contents	v
List of Figures	viii
List of Tables	xiv
Chapter 1: Introduction	1
Chapter 2: Review of Literature for MRP and P	16
RNase MRP	16
Protein Moiety Composition	16
Mitochondrial Activity	18
Nuclear Function	19
RNase P	20
Prokaryotic P	20
Mitochondrial P	22
Chloroplast P	23
Eukaryotic (nuclear) P	24
Evidence for the evolutionary relatedness between MRP and P	25
Chapter 3: Finding distantly related pRNA-like sequences in the chloroplast DNA of four green plant species.	
Introduction	27
Materials and Methods	28
Results	30
Discussion	40
Chapter 4: Evaluation of RNA biological secondary structure for use in determining evolutionary relationships.	
Preface	43
Abstract	45
Introduction	46
Materials and Methods	50
Results	51
Discussion	55
References	59

Chapter 5: Evaluation of folding programs for the analysis of evolutionary relationships of catalytic RNA molecules.	
Introduction	75
Materials and Methods	76
Results	79
Discussion	92
Chapter 6: Investigation of AT content and length on the comparison of folded pRNA sequences.	
Introduction	114
Materials and Methods	115
Results	115
Discussion	118
Chapter 7: Investigation of the percentage of pairing between nucleotides in folded secondary structures.	
Introduction	126
Materials and Methods	127
Results	129
Discussion	141
Chapter 8: Conclusions and Future Considerations.	
Evolution of mrpRNA	145
Comparison of biological secondary structures	146
Thermodynamic folding algorithms	147
The putative maize chloroplast pRNA	148
References:	149
Appendix 1: RNA secondary structures	
A: Biological RNA secondary structures of mrpRNA	160
B: Biological RNA secondary structures of pRNA	162
C: RNAstructure (Mfold) RNA secondary structures of mrpRNA	166
D: RNAstructure (Mfold) RNA secondary structures of pRNA	170
E: RNAdraw (RNAfold) RNA secondary structures of mrpRNA	180
F: RNAdraw (RNAfold) RNA secondary structures of pRNA	187
Appendix 2: Bracket Notation	198
Appendix 3: Input matrices for Neighbor	200

Appendix 4: Computer Program Parameters	201
Divide and Conquer	201
Dialign	202
ClustalX	203
Phylip package (DNAdist and Neighbor)	203
The Vienna RNA package	204
TreeView (Win32) (v1.40)	205
RNAstructure and Mfold	206
RNAdraw V1.1b	207
Sifold	208
Rsnfold	209
Pairs	209
Search from the FASTA package	211
Cl2bracket	212

List of Figures

Figure 1.1: Cartoon representation and biological secondary structure diagrams of <i>E. coli</i> and Human nuclear pRNA and human mrpRNA	5
Figure 1.2: Simplified secondary structure of mrpRNA showing features similar to that of eukaryotic, bacterial and mitochondrial pRNA.	6
Figure 1.3: Phylogenetic distribution of MRP and P.	11
Figure 1.4: Human pRNA biological and folded secondary structures.	13
Figure 1.5: <i>E. coli</i> pRNA biological and folded secondary structures.	14
Figure 1.6: Human mrpRNA biological and folded secondary structures.	15
Figure 2.1: <i>In vitro</i> processing of pre-rRNA showing the A2 and A3 cleavage sites.	19
Figure 3.1: Ssearch output from search of the maize chloroplast genome with the <i>Synechocystis</i> pRNA sequence.	31
Figure 3.2: ClustalX sequence alignments of the putative maize pRNA with A: <i>Synechocystis</i> pRNA and B: <i>Porphyra purpurea</i> chloroplast pRNA.	32
Figure 3.3: ClustalX sequence alignment of all four green plant chloroplast pRNAs.	33
Figure 3.4: ClustalX multiple sequence alignment of the four green plant chloroplast pRNA sequences and the <i>Synechocystis</i> pRNA sequence.	34
Figure 3.5: ClustalX multiple sequence alignment of the four green plant chloroplast pRNA sequences and the <i>Porphyra purpurea</i> chloroplast pRNA sequence.	35
Figure 3.6: The position of the green plant chloroplast pRNA (RNase P-like) sequences within the four chloroplast genomes.	36
Figure 3.7: Hypothetical secondary structures of the putative green plant chloroplast pRNA sequences from A: Maize, B: Rice, C: Tobacco and D: Spinach.	37

- Figure 3.8:** Structures folded using RNAstructure (mfold algorithm) of **A:** the putative maize chloroplast pRNA and **B:** *Synechocystis* pRNA showing both the 'fork' and 'can opener' motifs that have been found in other pRNA folded structures. 38
- Figure 3.9:** Structures folded using RNAdraw (RNAfold algorithm) of **A:** *Synechocystis* pRNA, **B:** the putative maize chloroplast pRNA and **C:** the *Porphyra purpurea* chloroplast pRNA, showing both the 'fork' and 'can opener' motifs found in other pRNA. 39
- Figure 4.1:** Comparison of three methods of constructing trees. **A:** Neighbor-joining with taxa loaded *A, B, C, D, E*; **B:** Neighbor-joining with taxa loaded *C, A, B, D, G, F, E*; **C:** Splitstree and **D:** refined Buneman. 65
- Figure 4.2:** **A:** Subtree of 16S rRNA bacterial and archaeobacterial sequences from the Ribosomal Database Project. **B:** Neighbor-joining tree of 16S rRNA Domain I length data. 66
- Figure 4.3:** Refined Buneman tree of mrpRNA sequences aligned by Divide and Conquer. 67
- Figure 4.4:** Refined Buneman trees of mrpRNA secondary structures compared by RNAdistance. 68
- Figure 4.5:** Refined Buneman tree of mrpRNA and pRNA sequences aligned by Divide and Conquer. 69
- Figure 4.6:** Refined Buneman tree constructed from pRNA and mrpRNA secondary structures. 70
- Figure 4.7:** Figure 4.8: Data used in 16S rRNA secondary structure analysis. **A:** Domain I of the 16S rRNA divided into areas. Numbering of the divisions is based on the *E. coli* 16S rRNA secondary structure. **B:** Species of bacteria and archaeobacteria used in this study their reference codes in the Ribosomal Database Project (RDP) and **C:** Matrix of differences between the areas in Domain I of the 16S rRNA secondary structure. 71
- Figure 4.8:** Data used in 16S rRNA Domain III and combined Domain I and III secondary structure analysis. **A:** Domain III of the 16S rRNA divided into areas. Numbering of the divisions is based on the *E. coli* 16S rRNA secondary structure. **B:** Matrix of differences between the areas of Domain III. **C:** Matrix of differences between the combined areas of Domains I and III. **D:** Neighbor-joining tree of Domain III lengths. **E:** Neighbor-joining tree of combined Domains I and III lengths. 72

- Figure 4.9:** Neighbor-joining of mrpRNA secondary structures compared by RNAdistance. 73
- Figure 4.10:** Neighbor-joining tree of MRP and pRNA sequences aligned by Divide and Conquer. 73
- Figure 4.11:** Neighbor-joining tree constructed from pRNA and mrpRNA secondary structures. 74
- Figure 5.1:** Human mrpRNA folded with RNAdraw to show **A:** uncorrected circular structure and **B:** 5' – 3' corrected structure. 77
- Figure 5.2:** pRNA sequences: **A:** Subtree of 16S rRNA sequences. Neighbor-joining trees of full structure format (f) of **B:** biological secondary structures. **C:** RNAstructure and RNAdraw folded secondary structures **D:** uncorrected, **E:** corrected. 80
- Figure 5.3:** Neighbor-joining tree of pRNA structures compared in the HIT structure format (h): **A:** biological secondary structures. **B:** folded by RNAstructure, and folded by RNAdraw **C:** uncorrected and **D:** corrected. 82
- Figure 5.4:** : Neighbor-joining tree of pRNA structures compared in the Weighted coarse format (w): **A:** biological secondary structures. **B:** folded by RNAstructure, and folded by RNAdraw **C:** uncorrected and **D:** corrected. 83
- Figure 5.5:** Neighbor-joining tree of pRNA structures compared in the Coarse format (c): **A:** biological secondary structures. **B:** folded by RNAstructure, and folded by RNAdraw **C:** uncorrected and **D:** corrected. 84
- Figure 5.6:** Neighbor-joining trees of mrpRNA **A:** aligned sequences. **B:** biological secondary structures: sequences folded by RNAstructure **C:** uncorrected and **D:** corrected. sequences folded by RNAdraw **E:** uncorrected and **F:** corrected. All structures are compared in the full (f) format. 86
- Figure 5.7:** Neighbor-joining trees of mrpRNA: **A:** biological secondary structures: sequences folded by RNAstructure **B:** uncorrected and **C:** corrected. sequences folded by RNAdraw **D:** uncorrected and **E:** corrected. All structures are compared in the HIT (h) format. 87

- Figure 3.8:** Structures folded using RNAstructure (mfold algorithm) of **A:** the putative maize chloroplast pRNA and **B:** *Synechocystis* pRNA showing both the 'fork' and 'can opener' motifs that have been found in other pRNA folded structures. 38
- Figure 3.9:** Structures folded using RNAdraw (RNAfold algorithm) of **A:** *Synechocystis* pRNA, **B:** the putative maize chloroplast pRNA and **C:** the *Porphyra purpurea* chloroplast pRNA, showing both the 'fork' and 'can opener' motifs found in other pRNA. 39
- Figure 4.1:** Comparison of three methods of constructing trees. **A:** Neighbor-joining with taxa loaded *A, B, C, D, E*; **B:** Neighbor-joining with taxa loaded *C, A, B, D, G, F, E*; **C:** Splitstree and **D:** refined Buneman. 65
- Figure 4.2:** **A:** Subtree of 16S rRNA bacterial and archaeobacterial sequences from the Ribosomal Database Project. **B:** Neighbor-joining tree of 16S rRNA Domain I length data. 66
- Figure 4.3:** Refined Buneman tree of mrpRNA sequences aligned by Divide and Conquer. 67
- Figure 4.4:** Refined Buneman trees of mrpRNA secondary structures compared by RNAdistance. 68
- Figure 4.5:** Refined Buneman tree of mrpRNA and pRNA sequences aligned by Divide and Conquer. 69
- Figure 4.6:** Refined Buneman tree constructed from pRNA and mrpRNA secondary structures. 70
- Figure 4.7:** Figure 4.8: Data used in 16S rRNA secondary structure analysis. **A:** Domain I of the 16S rRNA divided into areas. Numbering of the divisions is based on the *E. coli* 16S rRNA secondary structure. **B:** Species of bacteria and archaeobacteria used in this study their reference codes in the Ribosomal Database Project (RDP) and **C:** Matrix of differences between the areas in Domain I of the 16S rRNA secondary structure. 71
- Figure 4.8:** Data used in 16S rRNA Domain III and combined Domain I and III secondary structure analysis. **A:** Domain III of the 16S rRNA divided into areas. Numbering of the divisions is based on the *E. coli* 16S rRNA secondary structure. **B:** Matrix of differences between the areas of Domain III. **C:** Matrix of differences between the combined areas of Domains I and III. **D:** Neighbor-joining tree of Domain III lengths. **E:** Neighbor-joining tree of combined Domains I and III lengths. 72

- Figure 4.9:** Neighbor-joining of mrpRNA secondary structures compared by RNAdistance. 73
- Figure 4.10:** Neighbor-joining tree of MRP and pRNA sequences aligned by Divide and Conquer. 73
- Figure 4.11:** Neighbor-joining tree constructed from pRNA and mrpRNA secondary structures. 74
- Figure 5.1:** Human mrpRNA folded with RNAdraw to show **A:** uncorrected circular structure and **B:** 5' – 3' corrected structure. 77
- Figure 5.2:** pRNA sequences: **A:** Subtree of 16S rRNA sequences. Neighbor-joining trees of full structure format (f) of **B:** biological secondary structures, **C:** RNAstructure and RNAdraw folded secondary structures **D:** uncorrected, **E:** corrected. 80
- Figure 5.3:** Neighbor-joining tree of pRNA structures compared in the HIT structure format (h); **A:** biological secondary structures. **B:** folded by RNAstructure, and folded by RNAdraw **C:** uncorrected and **D:** corrected. 82
- Figure 5.4:** : Neighbor-joining tree of pRNA structures compared in the Weighted coarse format (w); **A:** biological secondary structures. **B:** folded by RNAstructure, and folded by RNAdraw **C:** uncorrected and **D:** corrected. 83
- Figure 5.5:** Neighbor-joining tree of pRNA structures compared in the Coarse format (c); **A:** biological secondary structures. **B:** folded by RNAstructure, and folded by RNAdraw **C:** uncorrected and **D:** corrected. 84
- Figure 5.6:** Neighbor-joining trees of mrpRNA **A:** aligned sequences, **B:** biological secondary structures; sequences folded by RNAstructure **C:** uncorrected and **D:** corrected; sequences folded by RNAdraw **E:** uncorrected and **F:** corrected. All structures are compared in the full (f) format. 86
- Figure 5.7:** Neighbor-joining trees of mrpRNA: **A:** biological secondary structures; sequences folded by RNAstructure **B:** uncorrected and **C:** corrected; sequences folded by RNAdraw **D:** uncorrected and **E:** corrected. All structures are compared in the HIT (h) format. 87

- Figure 5.8:** Neighbor-joining trees of mrpRNA: **A** biological secondary structures; sequences folded by RNAstructure **B**: uncorrected and **C**: corrected; sequences folded by RNAdraw **D**: uncorrected and **E**: corrected. All structures are compared in the Weighted Coarse (w) format. 88
- Figure 5.9:** Neighbor-joining trees of mrpRNA: **A** biological secondary structures; sequences folded by RNAstructure **B**: uncorrected and **C**: corrected; sequences folded by RNAdraw **D**: uncorrected and **E**: corrected. All structures are compared in the Coarse (c) format. 89
- Figure 5.10:** Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAstructure. Full format (f) - uncorrected. 96
- Figure 5.11:** Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAstructure. Full format (f) - corrected. 97
- Figure 5.12:** Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAstructure. HIT format (h) – uncorrected. 98
- Figure 5.13:** : Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAstructure. HIT format (h) – corrected. 99
- Figure 5.14:** Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAstructure. Weighted coarse (w) - uncorrected. 100
- Figure 5.15:** Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAstructure. Weighted coarse (w) - corrected. 101
- Figure 5.16:** Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAstructure. Coarse structure (c) – uncorrected. 102
- Figure 5.17:** Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAstructure. Coarse structure (c) – corrected. 103
- Figure 5.18:** Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAdraw. Full structure (f) – uncorrected. 104
- Figure 5.19:** Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAdraw. Full structure (f) – corrected. 105

Figure 5.20: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAdraw. . HIT structure (h) - uncorrected.	106
Figure 5.21: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAdraw. HIT structure (h) - corrected.	107
Figure 5.22: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAdraw. Weighted Coarse structure - uncorrected.	108
Figure 5.23: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAdraw. Weighted Coarse structure - corrected.	109
Figure 5.24: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAdraw. Coarse structure - uncorrected.	110
Figure 5.25: Neighbor-joining tree of mrpRNA and pRNA sequences folded by RNAdraw. Coarse structure - corrected.	111
Figure 5.26: Neighbor-joining tree of bacterial and organellar pRNA sequences folded by RNAdraw using full structure format. A: uncorrected and B: corrected.	112
Figure 5.27: Neighbor-joining tree of bacterial and organellar pRNA sequences folded by RNAstructure using full structure format. A: uncorrected and B: corrected.	113
Figure 6.1: Neighbor-joining tree of pRNA sequences compared to <i>E. coli</i> random sequences folded by RNAfold.	120
Figure 6.2: Neighbor-joining tree of pRNA sequences compared to <i>Porphyra</i> random sequences folded by RNAfold.	121
Figure 6.3: Neighbor-joining tree of pRNA sequences compared to putative maize chloroplast pRNA random sequences folded by RNAfold.	122
Figure 6.4: Neighbor-joining tree of pRNA sequences compared to <i>E. coli</i> random sequences folded by Mfold.	123
Figure 6.5: Neighbor-joining tree of pRNA sequences compared to <i>Porphyra</i> random sequences folded by Mfold.	124

- Figure 6.6:** Neighbor-joining tree of pRNA sequences compared to putative maize chloroplast pRNA random sequences folded by Mfold. 125
- Figure 7.1:** % pairing of 100 random and folded with RNAfold **A:** *E. coli* pRNA and **B:** *S. cerevisiae* mitochondrial pRNA and **C:** *Reclinomonas* mitochondrial pRNA 126
- Figure 7.2:** % pairing of 100 random sequences shuffled and folded with RNAfold. **A:** *Porphyra* chloroplast pRNA. **B:** the putative maize chloroplast pRNA. 127
- Figure 7.3:** % pairing of 100 random sequences shuffled and folded with RNAfold. **A:** Human nuclear pRNA. **B:** *S. cerevisiae* nuclear pRNA. **C:** Zebrafish nuclear pRNA. 129
- Figure 7.4:** % pairing of 100 random sequences shuffled and folded with RNAfold. **A:** Human mrpRNA. **B:** *S. cerevisiae* mrpRNA. **C:** *Arabidopsis* mrpRNA. 131
- Figure 7.5:** % pairing of 100 random sequences shuffled and folded with RNAfold. **A:** *Porphyra* chloroplast 50S ribosomal protein *L21*. **B:** *Arabidopsis* mitochondrial NADH dehydrogenase subunit 4L *nad4l*. **C:** maize chloroplast ribosomal protein A14 *rps14* **D:** *Porphyra* chloroplast allophycocyanin gamma chain protein *apcD*. 132
- Figure 7.6:** % pairing of 100 random sequences shuffled and folded with RNAfold. **A:** *Bacillus* nitrite reductase subunit *nasBD*. **B:** *E. coli* *cr1* protein. **C:** *Reclinomonas* mitochondrial ribosomal protein S12 *rps12*, and **D:** *Anabaena* sp. Nitrogen fixation protein *nifX2*. 132
- Figure 7.7:** Scatter plot of the % pairing against % AT for the RNAfold secondary structures for mrpRNA, eukaryotic pRNA, organellar and bacterial pRNA and protein coding mRNA sequences. 133
- Figure 7.8:** Scatter plot of the % pairing against length for the RNAfold secondary structures for mrpRNA, eukaryotic pRNA, organellar and bacterial pRNA and protein coding RNA sequences. 134
- Figure 7.9:** Graph of AT% against % pairing for RNAfold random sequences. 136
- Figure 7.10:** Graph of length against % pairing for RNAfold random sequences. 137

List of Tables

Table 1.1: Summary of characteristics and simplified secondary structure diagrams of bacterial, eukaryotic and organellar P and MRP.	3,4
Table 1.2: pRNA, mrpRNA and 16S rRNA sequences and secondary structures used in this study showing length, accession details, A + T % and from where the secondary structures were obtained.	9
Table 3.1: Chloroplast genomes, bacterial , cyanelle and chloroplast pRNA sequences and the putative green plant chloroplast pRNA sequences isolated in this chapter.	29
Table 4.1: RNase P and RNase MRP RNA sequences used in this study showing length, Accession details, A+T % and from where the secondary structures were obtained.	64
Table 5.1: mrpRNA and pRNA secondary structures that gave a circular structure when folded with RNAstructure and RNAdraw.	77
Table 7.1: % pairing, AT contents and lengths of A: mrpRNA and pRNA B: protein sequences used in this chapter.	124
Table 7.2: % pairing, AT contents and lengths of the random sequences formed from mrpRNA, pRNA and protein-coding RNA.	125
Table 7.3: Regression analysis of AT% against % pairing.	139
Table 7.4: Regression analysis of length against % pairing.	140

Chapter 1

Introduction

Bioinformatics, a new and exciting field in the biological sciences, is a powerful tool in the investigation of evolutionary relationships. Bioinformatics is applied here to examine two themes. Firstly, RNA secondary structure data is shown to reflect evolutionary relationships where the times of divergence are so old that there is little confidence in sequence data. Secondly, this secondary structure data is combined with sequence and functional data to examine the evolution of RNase MRP (MRP), especially the possibility of it being part of the RNA world.

RNase P (P) is already thought to be part of the RNA world, an early stage in the evolution of life, where RNA was both catalytic and the holder of the genetic information (Jeffares et al. 1998). MRP is thought to be evolutionary related to P due to functional and secondary structure similarities, but due to its presence only in eukaryotes, has not previously been considered to be part of the RNA world. These ribonucleoproteins (consisting of a catalytic RNA and at least one protein subunit) have RNA components (pRNA and mrpRNA) with little sequence homology, resulting in sequence alignments that have not enough reliability to confidently examine their evolutionary relatedness (Sbisà et al. 1996).

P cleaves tRNA precursors to form the mature 5' ends of tRNA molecules with activity being found all cells tested (i.e. universally) including prokaryotes, eukaryotes and also in organelles. Prokaryotic P consists of an RNA strand, and a single protein subunit, whereas the P encoded in the nucleus of eukaryotes has several protein subunits (Pace and Smith 1990). Fungi such as *Saccharomyces cerevisiae* and *Aspergillus nidulans* have retained their mitochondrial -encoded pRNA whereas vertebrate mitochondria and the fission yeast *Schizosaccharomyces pombe* have lost their pRNA gene and use a nuclear-encoded product. In plants, mitochondrial pRNA activity has been shown (Marchfelder and Brennicke 1993), but to date no genes have been characterised.

The secondary structure of prokaryotic pRNA has been seen in the past to show characteristic features for different phylogenetic groups of pRNA (Pace and Brown 1995) and consensus structures have been drawn for these groups of eubacteria and archaeobacteria (Haas et al. 1996, Pace and Brown 1995). This is an indication that some features in the pRNA secondary structure are fixed and others variable. For the

purposes of this study, prokaryotic pRNA includes that from eubacteria mitochondria, and plastids (chloroplast and cyanelle). The pRNA from archaeobacteria is not covered at this time due to processing power and time considerations.

MRP (Mitochondrial Ribosomal Processing) has been found only in eukaryotes initially as an endoribonuclease that cleaves RNA primers for the initiation of mitochondrial DNA replication (Morrissey and Tollervey 1995). Subsequently a nuclear function in rRNA processing was identified, consistent with its predominant localisation to the nucleolus (Lygerou et al. 1996). MRP consists of an RNA moiety and multiple protein subunits with at least 7 of these, Pop1p (Morrissey and Tollervey 1995), Pop3p (Dichtl and Tollervey 1997) Pop4p (Chu et al. 1997), Pop5p, Pop6p, Pop7p and Pop8p (Chamberlain et al. 1998) proteins being shared with P in the yeast *Saccharomyces cerevisiae*. It is possible that these proteins have structural characteristics that allow them to interact with both mrpRNA and pRNA. mrpRNA secondary structures (Schmitt et al. 1993) have only been characterised for eight species and show great similarity with each other despite being from plant, yeast and vertebrate species. The nucleotide sequences of these mrpRNAs vary greatly in length and nucleotide composition, making alignment of all eight sequences difficult.

Characteristics of MRP, eubacterial, eukaryotic and organellar P are summarised in Table 1.1. Cartoon representations and biological secondary structures of pRNA and mrpRNA show the sharing of some proteins between mrpRNA and the eukaryotic pRNA and the conserved presence of the pseudoknot pairing regions (Figure 1.1).

Comparisons of the RNA secondary structures between mrpRNA and pRNA have shown similarity in shape, especially in the 'cage region' of the RNA molecule in which there is the characteristic pseudoknot formation (Forster and Altman 1990). (Pseudoknots are structural elements that may act as a recognition site for proteins involved in replication initiation or translational regulation. The NMR structure of the classical pseudoknot has been determined (Kolk et al. 1998).) However, to date, there has been no published quantitative comparison of pRNA and mrpRNA secondary structure. When pRNA and mrpRNA secondary structures are broken down into simplified structures it can be seen that a large proportion of the secondary structure is shared between these two RNA molecules (Figure 1.2).

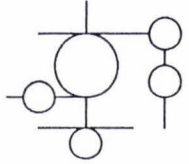
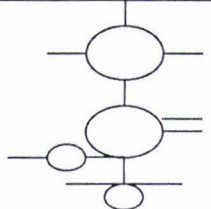
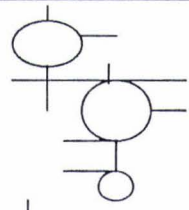
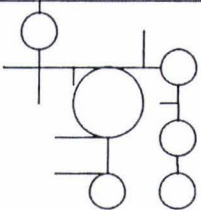
	Activity	Complex	Reaction Catalysed	Encoded	RNA transport	RNA structure (Simplified)	Comments
RNase MRP	Nucleolus, Mitochondria	RNA + Protein More than one protein subunit. In <i>S. cerevisiae</i> POP-1 and SMN1 identified.	rRNA processing in nucleolus, Cleaves RNA primers in mitochondria.	Nucleus	To the nucleolus. To the mitochondria.		Pop-1, Pop3 and Pop4 proteins shared with <i>S. cerevisiae</i> P. SMN1 unique to MRP.
Eukaryotic RNase P	Nucleus (Can also be found in mitochondria and chloroplasts)	RNA + Protein More than one protein subunit involved.	Cleaves pre-tRNAs to form mature tRNAs	Nucleus	To the mitochondria. Stays within the nucleus.		Many mammalian pRNA sequences in the databases, but none from the 'lower' eukaryotes and the amitochondrial eukaryotes as yet.
Eubacterial RNase P	Cell	RNA + Protein (RNA can be catalytic on its own). One protein in the complex.	Cleaves pre-tRNAs to form mature tRNAs	Chromosome	Within the cell.		pRNA's from many eubacterial species isolated but only <i>E. coli</i> RNase P studied in detail.
Mitochondrial RNase P	Mitochondria	RNA + Protein Protein is nuclear encoded. Unsure of how many subunits involved	Cleaves pre-tRNAs to form mature tRNAs	Vertebrates and <i>S. pombe</i> use nuclear encoded gene. Other yeasts, plants and <i>Reclinomonas americana</i> encode a mitochondrial gene.	To Mitochondria. Within the mitochondria.		RNA structure is much like that of the bacterium <i>Rhodospirillum</i> . Is highly A-U rich and very variable in size.

Table 1.1: Summary of characteristics and simplified secondary structure diagrams of MRP, eubacterial, eukaryotic and organellar P.

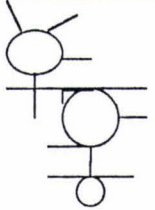
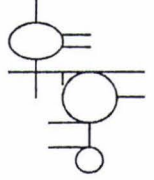
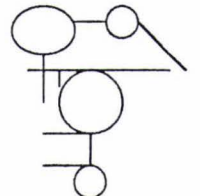
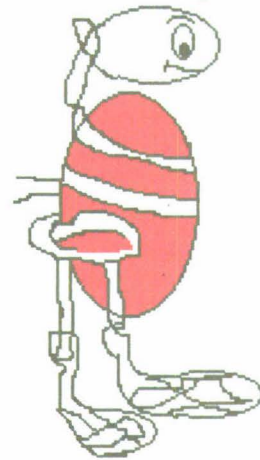
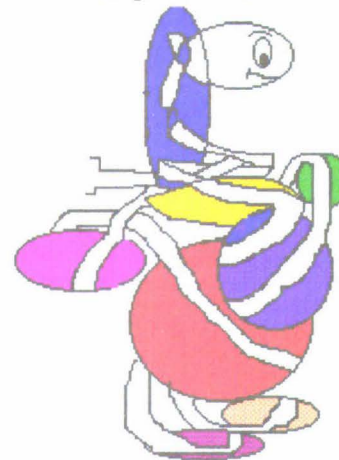
Activity	Where the Activity is found	Complex	Reaction Catalysed	RNA encoded	RNA localisation	RNA structure (Simplified)	Comments
Chloroplast RNase P	Chloroplast	RNA + protein Chloroplast RNA from <i>Porphyra purpurea</i> has been sequenced.	Cleaves pre-tRNAs to form mature tRNAs	Chloroplast	Chloroplast		Only sequence found so far is in the <i>Porphyra purpurea</i> chloroplast.
Cyanelle RNase P (Cyanophora paradoxa)	Cyanelle	RNA + Protein Thought to be one protein in complex. Eubacterial-like.	Cleaves pre-tRNAs to form mature tRNAs	Cyanelle genome.	Within the cyanelle.		RNA structure is very similar to that of the cyanobacteria.
Archaeal RNase P	Cell	RNA + Protein Thought to be one protein in complex. Eubacterial-like.	Cleaves pre-tRNAs to form mature tRNAs	Cell	Within the cell.		RNA is considered eubacterial-like.

Table 1.1 continued: Summary of characteristics and simplified secondary structure diagrams of MRP, eubacterial, eukaryotic and organellar P.

Bacterial RNase P



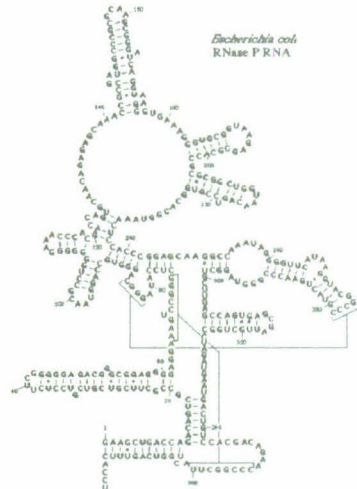
Eukaryotic RNase P



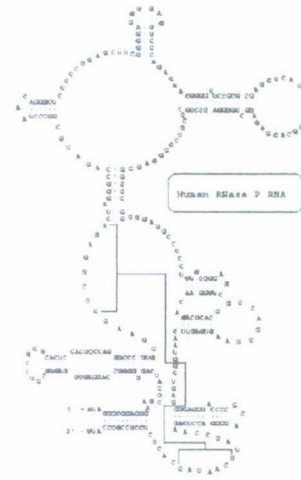
RNase MRP



A



B



C

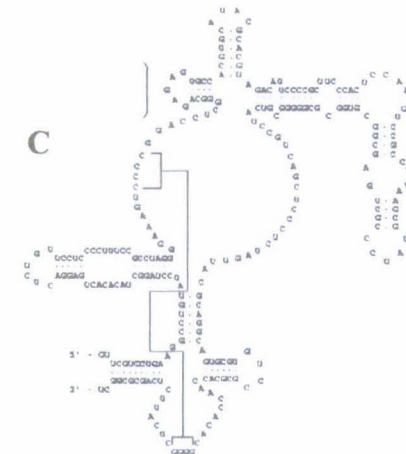
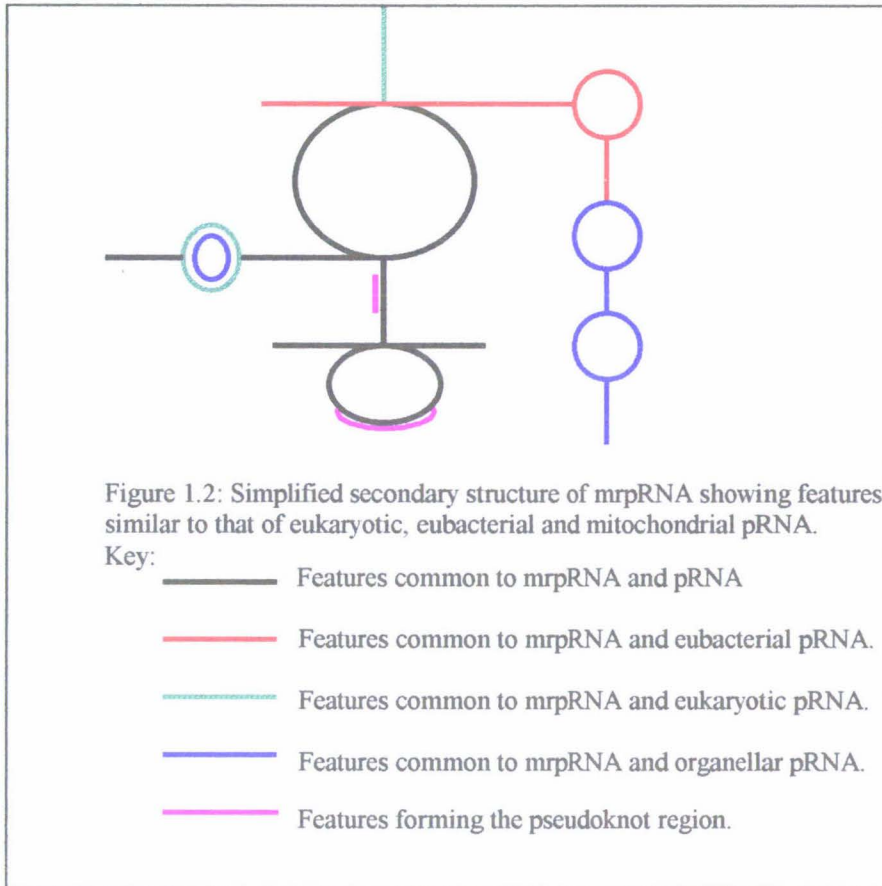


Figure 1.1. Cartoon Representation and RNA biological secondary structure diagrams.

A: *E. coli* pRNA (From the RNase P Database, Brown 1998) associating with one protein, **B:** Human nuclear pRNA (Redrawn from Altman et al. 1993) associating with multiple proteins, **C:** Human RNase mrpRNA (Redrawn from Schmitt et al. 1993) also associating with multiple proteins, some of which are shared with P. The solid lines indicate the long range pairing important in tertiary structure formation (including the Pseudoknot formation).



The secondary structure and functional similarities between MRP and P have led to the conclusion that these two ribonucleoproteins (RNP's) are evolutionary related (Morrissey and Tollervey 1995). Both the P and MRP ribozymes cleave RNA's to generate 5' phosphate and 3' hydroxyl termini in a reaction requiring divalent cations (Forster and Altman 1990). They are both sensitive to puromycin, an antibiotic which inhibits pre-tRNA processing (Potuschak et al. 1993), and enzymatic activities from P and MRP isolated from several organisms cofractionate through multiple stages of biochemical purification (Paluh and Clayton 1995). It has been reported that MRP and P may be involved together in a macromolecular complex within the nucleolus (Lee et al. 1996). A contrary theory, however, is that the relationship between MRP and P may be of a functional nature based on their sharing of many protein subunits (Sbisà et al. 1996).

This study investigated three general hypotheses, based on functional characteristics, of the relatedness of P and MRP. pRNA, mrpRNA and 16S rRNA sequences and secondary structures used in this study, are shown in Table 1.2.

The three groups of hypotheses are as follows:

I MRP evolved from an eukaryotic nuclear P in the nucleus of the eukaryotic cell. This could occur by gene duplication followed by divergence of function of the two homologues. This is the theory most commonly suggested in previous studies (Morrissey and Tollervey 1995, Reddy and Shimba 1996, Chamberlain et al. 1996). MRP would have been incorporated into multiple eukaryotic functions and has also gained an essential function in mitochondria. Under this hypothesis MRP is found only in eukaryotes because it was never in any of the other lineages! MRP is present in animals, yeasts, and plants indicating an early divergence from P; however, MRP need not have been present in all early eukaryotes. We would expect under this hypothesis the secondary structures of the mrpRNA to be more similar to eukaryotic pRNA than to prokaryotic pRNA.

Under this hypothesis MRP is an exception to the transfer process of catalysis (RNA to RNP to protein) (Jeffares et al. 1998) with a ribonucleoprotein taking on a new catalytic function after the widespread availability of protein catalysts.

II MRP evolved from an endosymbiont P. MRP could have evolved from the hypothetical endosymbiotic fusion that formed the first eukaryote (Gupta and Golding 1996) or by some later endosymbiosis that led to the mitochondrion. The endosymbiotic origin theory accounts for the essential mitochondrial function of MRP. It has been shown that organellar DNA can be transferred to the nucleus and yet retain a function in the organelle (Brennicke et al. 1993, Wischmann and Schuster 1995, Blanchard and Schmidt 1995). This theory proposes that MRP picked up the additional rRNA processing functions in the nucleus. We might expect here that mrpRNA would retain some organellar characteristics such as a higher A + T content in nucleotide sequence and be more closely related in secondary structure to that of the organellar or prokaryotic pRNA.

III MRP and P evolved in the RNA world. The RNA world hypothesis suggests that DNA and proteins evolved from a world in which RNA was the both the catalytic and information storage molecule, and that today's catalytic RNA species are molecular relics from this time. There are three main criteria used to evaluate the

antiquity of an RNA molecule (Jeffares et al. 1998) and pRNA fits all three of these criteria by being ubiquitous, catalytic and central to metabolism. MRP on the other hand fits only the last two criteria, being present only in the eukaryotic lineage. A central concept to the RNA world is that proteins with superior catalytic properties have gradually replaced RNA as the catalytic molecule (and that no novel catalytic RNAs would be formed after the advent of efficient protein synthesis, Jeffares et al. 1998).

However, it is difficult to see how a molecule such as MRP could have evolved only in the eukaryotic lineage and then integrate itself so intimately into rRNA processing, mitochondrial genome replication, and perhaps other functions central to eukaryotic metabolism. It has been found that eukaryotes carry more proposed 'relics' of the RNA world than prokaryotes. These 'relics' include small nucleolar RNAs, spliceosomes, telomerase, and self-splicing introns, which are all absent from prokaryotes (Jeffares et al 1998). MRP was the only widely occurring catalytic RNA not suggested to be a relic from the RNA world in Jeffares et al. 1998.

Again there are several variants of this hypothesis; MRP could have evolved from P, P evolving from MRP, and MRP and P evolving independently in the RNA world.

With such an early divergence expected between pRNA and mrpRNA (at least back to the divergence of eukaryotes), nucleotide sequence alignments may not be reliable enough to determine with confidence any evolutionary relationship. It is expected, however, that examination of the RNA secondary structure may yield the required information when the sequence data cannot.

It has been shown that many sequences can fit the same secondary structure (Fontana et al. 1993) which allows the catalytic RNA sequence to vary even if the function of the molecule remains unchanged. The secondary structure of the catalytic RNA molecule has both fixed 'motifs' that represent areas that are critical to maintaining the function, and other regions that are free to vary in presence or size. It is expected that these fixed and variable regions of the catalytic RNA secondary structure will change according to the evolution of the function of the molecule, and thus may be used to determine evolutionary relationships when the sequence data may not. Quantitative comparisons of pRNA and mrpRNA secondary structures are used here to calculate distances between these molecules in order to assess their relatedness.

	Accession Number	Length of Sequence	A + T %	Secondary Structure Reference
pRNA Sequences				
Eubacterial pRNA				
Synechocystis sp. PCC6803	X65707	437	48	P
Anabaena sp. PCC 7120	X65648	465	47	P
Anacystis nidulans PCC6301	X63566	385	43	P
Pseudoanabaena sp. PCC 6903	X73135	450	52	P
Escherichia coli	M17569	377	38	P
Bacillus subtilis	M13175	401	51	P
Rhodospirillum rubrum	M59355	429	29	P
Agrobacterium tumefaciens	M59354	402	36	P
Mitochondrial pRNA				
Reclinomonas americana mitochondria	AF007261	312	75	P
Saccharomyces cerevisiae mitochondria	U46121	448	87	No structure
Aspergillus nidulans mitochondria	X93307	300	81	No structure
Plastid pRNA				
Porphyra purpurea chloroplast	U38804	383	63	P
Cyanophora paradoxa Cyanelle	X89853	350	67	P
Eukaryotic pRNA				
Human (nuclear)	X15624	340	36	Altman et al. 1993
Mouse (nuclear)	L08802	288	33	Altman et al. 1993
Danio rerio (nuclear) Zebrafish	U50408	308	43	No structure
Saccharomyces cerevisiae (nuclear)	M27035	368	48	Tranguch and Engelke 1993
Schizosaccharomyces pombe (nuclear)	X04013	373	48	Tranguch and Engelke 1993
mrpRNA Sequences				
Human	X51867	264	36	Schmitt et al. 1993
Bovine	Z25280	277	39	Schmitt et al. 1993
Mouse	J03151	275	36	Schmitt et al. 1993
Rat	J05014	273	35	Schmitt et al. 1993
Xenopus (frog)	Z11844	277	45	Schmitt et al. 1993
Arabidopsis thaliana	X65942	260	49	Kiss et al. 1992
Saccharomyces cerevisiae	Z14231	339	60	Kiss et al. 1992
Schizosaccharomyces pombe	X04013	399	57	Paluh and Clayton 1995
16S rRNA structures				
	RDP sequence			
				RDP
Escherichia coli	E.coli	-	-	RDP
Clostridium innocuum	C.innocuum	-	-	RDP
Methanococcus vannielli	Mc.vanniell	-	-	RDP
Frankia sp.	Fra.spORS	-	-	RDP
Streptomyces coelicolor	Strn.coelic	-	-	RDP
Thermus thermophilus	T.thermoph	-	-	RDP
Bacillus subtilis	B.subtilis	-	-	RDP
Agrobacterium tumefaciens	Ag.tumefac	-	-	RDP
Spirochaeta aurantia	Spi.aurant	-	-	RDP
Thermoplasma acidophilum	Tpl.acidop	-	-	RDP
Mycoplasma capricolum	M.capricol	-	-	RDP
Methanobacterium formicicum	Mb.formici	-	-	RDP
Pseudomonas testosteroni	Ps.testost	-	-	RDP

Table 1.2: pRNA , mrpRNA and 16S rRNA sequences and secondary structures used in this study showing length, accession details, A+T % and from where the secondary structures were obtained Key: P Obtained from the RNase P Database (Brown 1997).

RDP Obtained from the Ribosomal Database Project (Maidak et al. 1997).

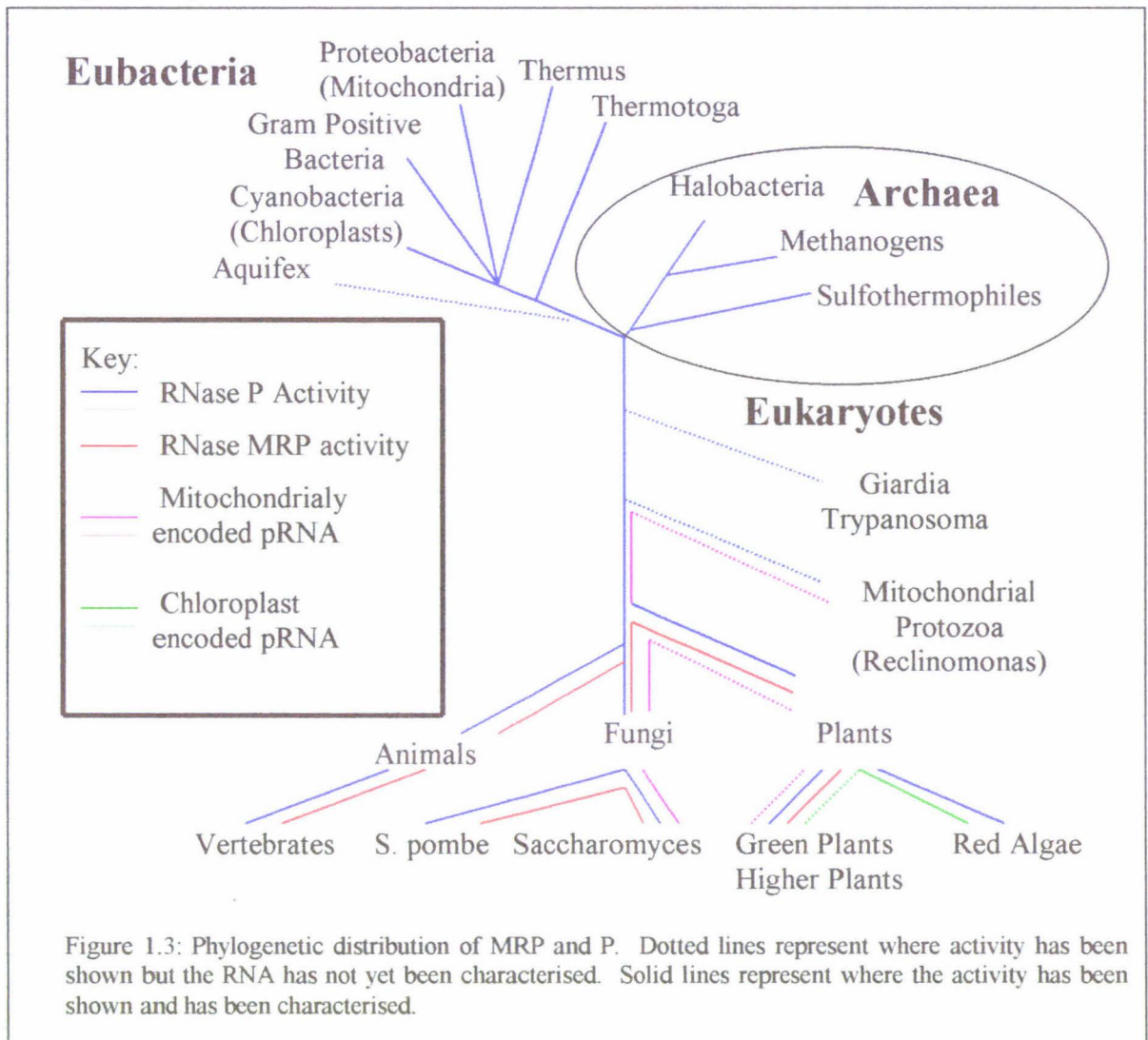
This study examined two types of RNA secondary structure. The first is the secondary structure that the RNA forms in nature and is referred to here as the "biological secondary structure". The biological secondary structures of eubacterial pRNA have been studied extensively (Haas et al. 1994, Haas et al. 1996a, Haas et al. 1996b, Green et al. 1996) and consensus structures calculated. Eukaryotic and organellar pRNA biological secondary structures are not as well defined with published hypothetical structures being used here. Some organellar sequences used in this study do not have any published secondary structure and are only used when sequence data alone is required.

The second type of secondary structure is calculated from the nucleotide sequence data using folding programs. Such structures are determined only from the nucleotide sequence data and need not have any relationship to the function of the molecule. Thus, the calculated secondary structures may not have the same fixed and varied regions that are shown in the biological structures (Zuker 1989).

Within the fixed regions of the biological secondary structure it is expected that nucleotide changes in one part of a helix will be met by a corresponding change in another part of the sequence to allow the helix to remain unchanged. Thus it is still expected that sequences of similar functions will form similar secondary structures with the folding programs allowing the formation of a recognisable structural 'motif'. These motifs are possible identification features that could be used in the characterisation of putative catalytic RNA sequences. Secondary structures folded from pRNA and mrpRNA sequences with folding programs are examined for use in the characterisation of putative catalytic RNA sequences.

Organellar genomes (mitochondria and chloroplast) offer a unique opportunity for the testing of searching, gene identification, and characterisation techniques. These genomes are small and many have been completely sequenced, and are available in databases such as GenBank. However the high AT content of organellar genomes often makes them hard to search with standard searching algorithms. Searching databases with a sequence of high AT content gives a high background of non-relevant matches often obscuring meaningful results. The distribution of pRNA and mrpRNA (Figure 1.3) shows that although pRNA is found encoded in the mitochondrial DNA of plants, there is to date, no published green plant chloroplast-encoded pRNA sequences. To test the feasibility of using RNA secondary structure to

characterise potential pRNA sequences, green plant chloroplast genomes were searched for putative pRNA sequences.



It is only recently that pRNA was characterised from the chloroplast of the red alga *Porphyra purpurea* (Reith and Munholland 1995), and from the cyanelle (a chloroplast-like plastid that still retains a cell wall) of *Cyanophora paradoxa* (Baum et al. 1996). Although it is expected that sequence homology between known pRNA sequences and putative green plant chloroplast pRNA sequences would be low, it is still expected that secondary structure (both a theoretical biological structure based on other pRNA structures and a folded structure), would show identifying secondary structure characteristics. One of the putative green plant chloroplast pRNA sequences (from the *Zea mays* – maize chloroplast) is examined more fully with other pRNA and mrpRNA sequences in this study.

There is a possibility that folded structures could be used in the same way as the biological structures, for determining evolutionary relationships. The biological and folded structures from two folding programs are shown for Human pRNA (Figure 1.4), *Escherichia coli* pRNA (Figure 1.5), and Human mrpRNA (Figure 1.6). These figures highlight how different the calculated structures are from the biological structures but also the similarities between the structures formed by the two different folding programs.

Problems with the use of folding programs in the analysis of catalytic RNA may include how much influence characteristics such as the AT content and sequence length, have on the estimated structure. These factors are examined here using random sequences derived by shuffling pRNA and mrpRNA sequences of varying length and AT content. Protein-coding RNA sequences are also used as controls in order to evaluate any trends that may be used in identifying putative catalytic RNA sequences. The amount of pairing that is present in a folded structure could also be another tool in the identification of catalytic RNA sequences.

In summary, this thesis looked at four main issues. The first was the evolution of MRP and its relationship to P. The second was the use of RNA secondary structure in the characterisation of putative pRNA sequences from chloroplasts. The third was the use of biological secondary structure in determining evolutionary relationships, and the fourth was the evaluation of the structural output from folding programs. The techniques developed here may, in future, be applied to other RNA molecules especially those associated with the RNA world as well as the analysis of newly discovered potential RNA molecules.

Human nuclear pRNA

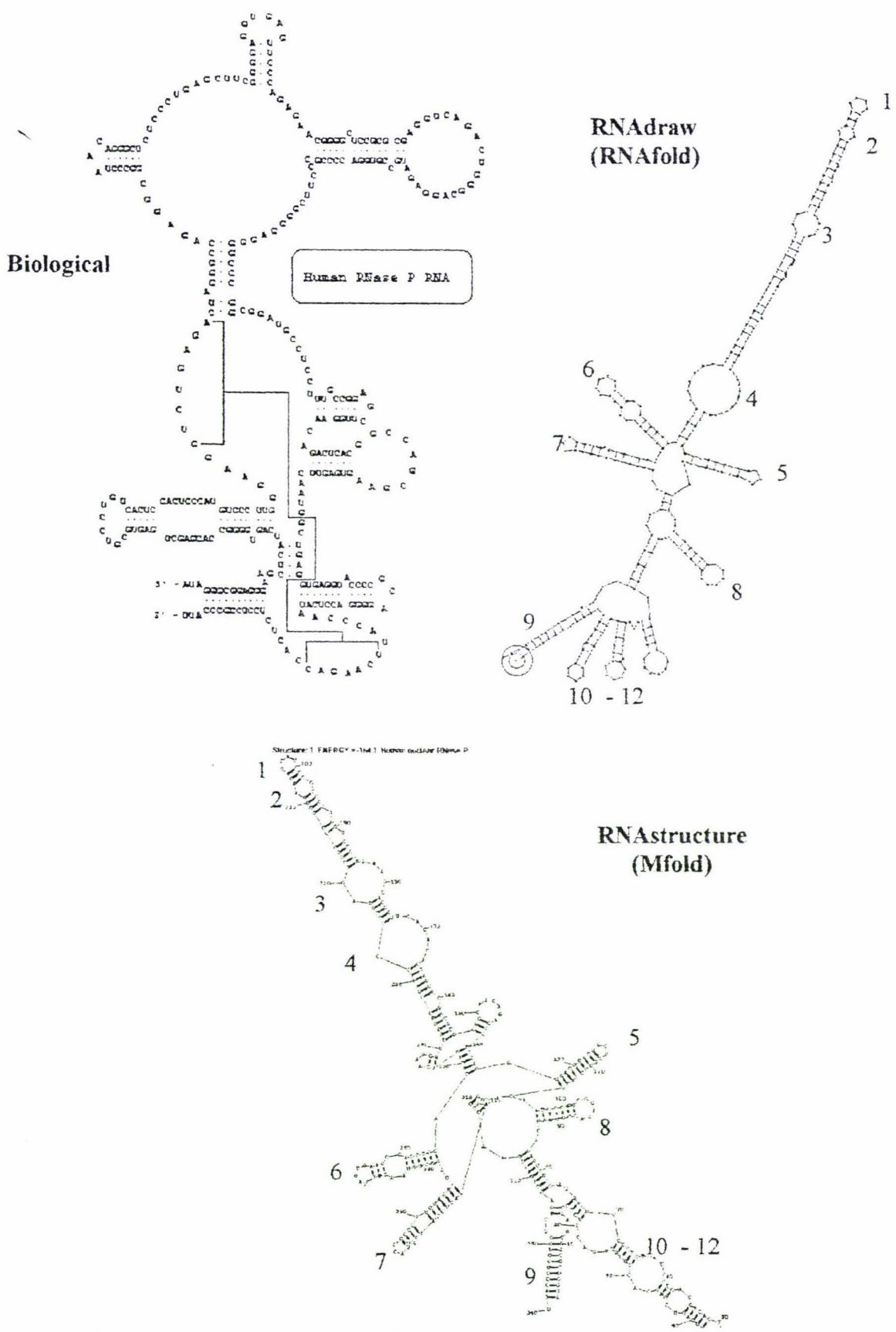
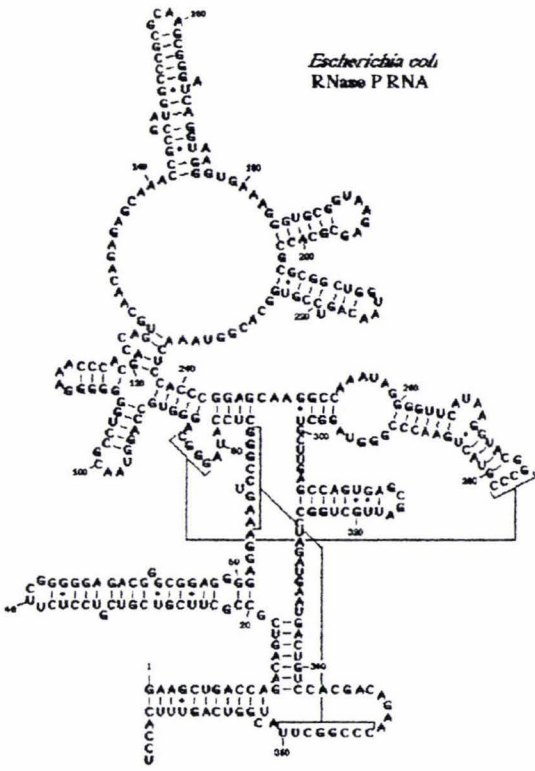


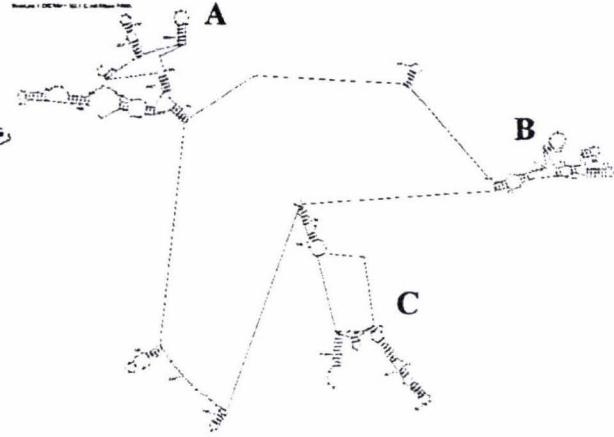
Figure 1.4: Human pRNA biological and folded secondary structures. Numbers 1 to 12 represent features that may be common to both the RNAstructure and the RNAdraw structures.

E. coli pRNA

Biological



RNAstructure
(Mfold)



RNAdraw
(RNAfold)

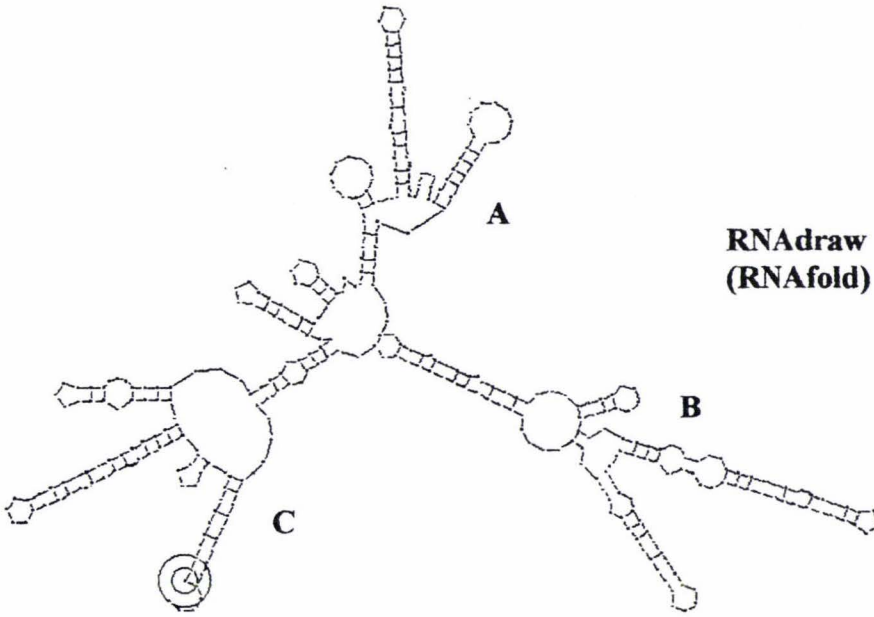


Figure 1.5: *E. coli* pRNA biological and folded secondary structures. A, B, and C represent features that may be common to both the RNAstructure and the RNAdraw structures.

Human mrpRNA

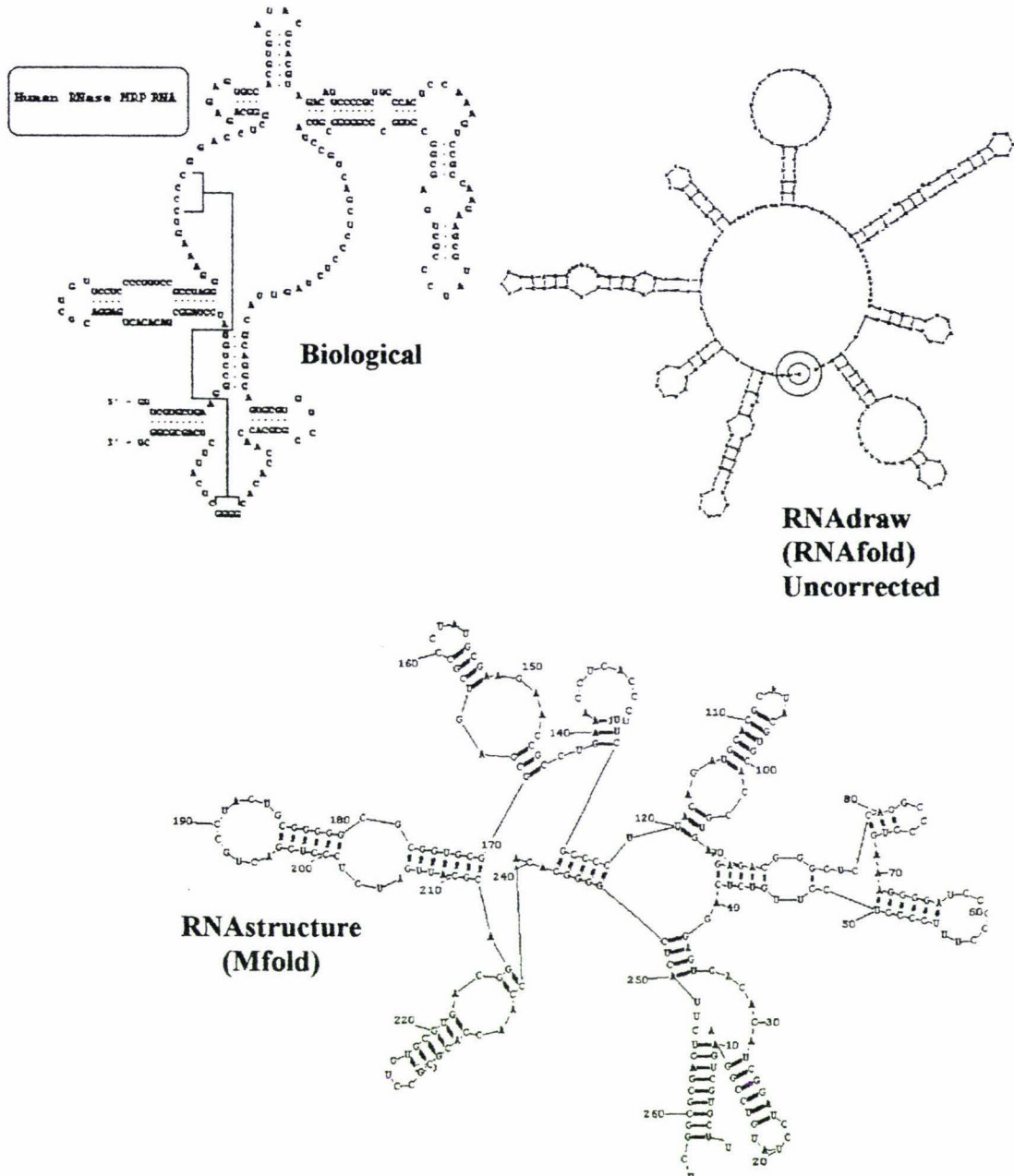


Figure 1.6: Human mrpRNA biological and folded secondary structures. The RNAdraw structure in this case has formed short range pairing in preference to the long range pairing required to pair the 5' and 3' ends. This structure is corrected when required to form 5' - 3' pairing.